Part-of-speech Proportion as an Index of Formality and Informality: The Case of Japanese

QUALICO 2023, at the University of Lausanne, 28-30 June 2023

Please leave your questions and/or comments here if any.

Or simply send an email to:

Tatsuhiko Matsushita (National Institute for Japanese Language and Linguistics, NINJAL)

tatsu.matsushita@ninjal.ac.jp

Background

- There are many studies that have applied part-of-speech (POS or word class) distributions to genre analysis (For Japanese, e.g., Kabashima, 1954). Some of them use the ratio of a particular POS as an index of formality.
- However, few studies, at least in Japanese, have analyzed the POS distributions of different types of text domains in a large, up-to-date corpus as an indicator of formality.
- **Research Question: Is there any pattern of POS distributions among different text domains of a large contemporary Japanese corpus? If yes, does it relate to the degree of formality?**

Design

- Corpus: Balanced Corpus of Contemporary Written Japanese (BCCWJ) 2009 monitor version (NINJAL, 2009) consisting of book texts and Internet forum sites with approximately 3.3 million lemmas
- Procedure

Step 1 : Dividing the entire corpus into **10 domains** based on the NDC (Nippon Decimal Classification: A book code system for Japanese libraries)
Step 2 : Analyzing the texts in these **10 domains** with the morphological analyzer MeCab Ver. 0.98 (Kudo, 2009) and the morphological dictionary UniDic Ver.1.3.11 (Den et al., 2009) with POS tagger
Step 3: Calculating the ratio of each POS in the total number of words
Step 4: Examining correlations between the POS ratios by domains

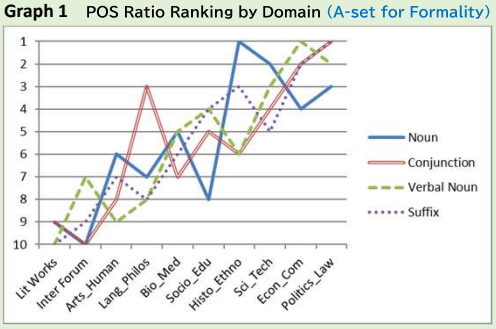
The 10 domains [# of lemmas in million]

1. **Lit Works** [8.25]
2. **Inter Forum**: Internet Q & A Forum. [5.22]
3. **Arts Human**: Arts and Other Humanities [3.02]
4. **Lang Philos**: Languages, Language and Philosophy [2.13]
5. **Bio Med**: Biology and Medicine [2.25]
6. **Socio Edu**: Sociology, Education and Other Social Issues [3.00]
7. **Histo Ethno**: History and Ethnology [3.34]
8. **Econ Com**: Economics and Commerce [2.21]
9. **Sci Tech**: Science and Technology [1.51]
10. **Politics Law** [1.88]

Results

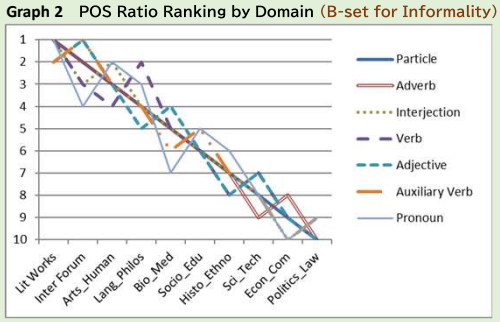
A-set POS (Formality): Conjunction, Verbal Noun, Suffix, Noun

Graph 1 POS Ratio Ranking by Domain (A-set for Formality)



B-set POS (Informality): Interjection, Pronoun, Adjective, Adverb, Auxiliary Verb, Verb, Particle

Graph 2 POS Ratio Ranking by Domain (B-set for Informality)



C-set POS (Non-indexical): pre-noun adjectivals, prefixes, and Nominal Adjectives

Graph 3 POS Ratio Ranking by Domain (Non-indexical POS)

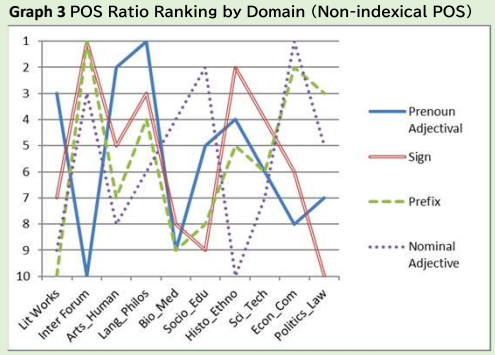


Table 1 Proportion of Indexical Sets of POS by the 10 Domains in BCCWJ 2009 (%)

POS \ Domain	Infromal ← ← ← ←						→ → → → Formal			
	Lit Works	Inter Forum	Arts Human	Lang Philos	Bio Med	Socio Edu	Histo Ethno	Econ Com	Sci Tech	Politics Law
Conjunction	0.3	0.2	0.4	0.6	0.5	0.5	0.5	0.5	0.6	0.6
Verbal Noun	3.2	5.5	4.7	5.5	6.5	7.8	5.6	8.2	9.5	9.3
Suffix	2.7	2.8	3.5	3.1	3.7	4.0	4.1	3.9	4.2	4.4
Noun	21.9	21.4	25.6	25.2	26.5	25.1	30.1	28.5	27.4	27.8
Subtotal (A-set)	28.1	29.9	34.2	34.4	37.2	37.5	40.3	41.1	41.6	42.2
Ratio Size	(A-set) Smaller ← ← ← ← (B-set) Larger ← ← ← ←						→ → → → Larger → → → → smaller			
Particle	33.3	32.2	31.9	31.7	31.1	31.0	30.1	29.6	29.6	29.4
Verb	15.5	14.3	14.3	14.9	14.3	14.2	13.2	13.3	13.2	13.1
Auxiliary Verb	12.6	14.3	10.4	10.2	9.5	9.5	9.2	8.9	8.4	8.6
Adverb	2.5	2.1	2.0	1.8	1.7	1.6	1.5	1.4	1.4	1.3
Adjective	2.0	2.2	1.7	1.5	1.7	1.4	1.2	1.3	1.2	1.0
Pronoun	2.5	1.6	1.9	1.8	1.2	1.3	1.3	1.0	1.0	1.0
Interjection	0.5	0.1	0.2	0.1	0.1	0.1	0.1	0.1	0.0	0.1
Subtotal (B-set)	68.8	66.9	62.4	61.9	59.5	59.1	56.6	55.5	54.8	54.4
(B-A+100)/2 (*)	29.6	31.5	35.9	36.2	38.8	39.2	41.9	42.8	43.4	43.9

***This formula is basically the same idea as the index *F* proposed by Heylighen & Dewaele (1999).**

- Red boxes in Table 1 show the domains grouped by **cluster** analysis based on the POS distribution.
- **Correlation** between **Subtotals (A-set) & (B-set)**: $r = -.999$, $p<.001$
- Noun has indexicality for **formality** in many languages, while interjection, adverb, and verb are for **informality**., since nouns carry **information transfer** and increase in proportion in **context-independent expressions**, whereas interjections and pronouns are **deictic and context-dependent** (e.g., Heylighen & Dewaele, 1999; Kabashima, 1954).
- Some other POS also have strong indexicality in Japanese. There will be **language-specific grammatical factors** as well.

Table 2 Correlation Coefficient (*r*) between the POS Ratios by the 10 Domains in BCCWJ 2009 (* p<0.05, ** p<0.01)

Negative Correlation

Indexicality	Informality							Non-indexical			Formality			
POS	Interjection	Pronoun	Adjective	Adverb	Auxiliary Verb	Verb	Particle	Prefix	Nominal Adjective	Prenoun Adjectival	Noun	Suffix	Verbal Noun	Conjunction
Interjection	1	.849**	0.5249	.803**	0.5128	.707*	.728*	.414	.547	.279	-.539	-.615	-.731*	-.542
Pronoun		1	.703*	.938**	.693*	.882**	.926**	.193	.477	.323	-.710*	-.841**	-.903**	-.596
Adjective			1	.896**	.923**	.765**	.889**	.102	.144	.322	-.873**	-.907**	-.742*	-.884**
Adverb				1	.854**	.885**	.976**	.183	.368	.023	-.849**	-.919**	-.883**	-.795**
AuxiliaryVerb					1	.662*	.828**	.098	.217	.431	-.883**	-.892**	-.698*	-.916**
Verb						1	.943**	.164	.149	.203	-.803**	-.867**	-.785**	-.502
Particle							1	.175	.285	.062	.862**	-.920**	-.881**	-.728*
Prefix								1	.442	.334	-.156	-.022	.363	.193
NominalAdjective									1	.359	-.109	.230	.669*	.315
PrenounAdjectival										1	.329	.089	-.267	.463
Noun											1	.853**	0.555	.745*
Suffix												1	.816**	.741*
VerbalNoun													1	.662*
Conjunction														1

B-set (Informality)

C-set (Non-indexical)

A-set (Formality)

Conclusion

★There is a clear pattern of POS distribution in Japanese texts, which seemingly shows the formality and Informality. Specific POS concerning formality will be partially different between languages.

This work was supported by JSPS KAKENHI Grant Number 20KK0005 and 23H00072.

*This study is extracted from and added to Chapter 4 of the presenter's doctoral thesis (Matsushita, 2012).

References (excerpt)

Den, Y., Yamada, A., Ogura, H., Koiso, H., & Ogiso, T. (2009). UniDic (digitized dictionary for morphological analysis) 1.3.11
Heylighen F. & Dewaele, J. (1999). Formality of language: definition, measurement and behavioral determinants. Internal Report, Center "Leo Apostel", Free University of Brussels. Downloaded from https://www.researchgate.net/profile/Francis-Heylighen/publication/2420048
Kabashima, T. (1954). Gendaibun-ni-okeru hinshi-no hititsu-to sono zougen-no youin-ni tsuite (On the ratio of parts of speech in present-day Japanese and the cause of its fluctuation). *Kokugogaku (Japanese Linguistics), 18*, 15-20.
Kudo, T. (2009). MeCab (morphological analyzer) 0.98.
Matsushita, T. (2012). In what order should learners learn Japanese vocabulary? A corpus-based approach (PhD thesis). Victoria University of Wellington. Downloaded from https://researcharchive.vuw.ac.nz/xmlui/handle/10063/4476
NINJAL (The National Institute for Japanese Language). (2009). Balanced Corpus of Contemporary Written Japanese (2009 monitor version). (Available by application at that time).